What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet

G. Eysenbach [1) 2)], Ch. Kohler [2)]

1) Centre for Global eHealth Innovation, University Health Network, Toronto General Hospital, Canada 2) Dept. of Clinical Social Medicine, Unit for Cybermedicine, University of Heidelberg, Germany

While health information is often said to be the most sought after information on the web, empirical data on the actual frequency of health-related searches on the web are missing. In the present study we aimed to determine the prevalence of health-related searches on the web by analyzing search terms entered by people into popular search engines. We also made some preliminary attempts in qualitatively describing and classifying these searches. Occasional difficulties in determining what constitutes a "health-related" search led us to propose and validate a simple method to automatically classify a search string as "health-related". This method is based on determining the proportion of pages on the web containing the search string and the word "health", as a proportion of the total number of pages with the search string alone. Using human codings as gold standard we plotted a ROC curve and determined empirically that if this "co-occurance rate" is larger than 35%, the search string can be said to be health-related (sensitivity: 85.2%, specificity 80.4%). The results of our "human" codings of search queries determined that about 4.5% of all searches are "health-related". We estimate that globally a minimum of 6.75 Million health-related searches are being conducted on the web every day, which is roughly the same number of searches that have been conducted on the NLM Medlars system in 1996 in a full year.

Introduction

It is often said that the most common reason for why people go online is to search for health information. This statement seems to be based primarily on survey research such as the Pew Internet Survey which claims that 55% of those with Internet access, have used the Web to get health or medical information [1]. However, it is unclear what the actual volume and prevalence of health-related searches on the web is in relation to the total number of searches conducted daily on the Internet. Given the rich data source the Internet represents to study personal health information seeking behavior there is a surprising dearth of evidence on what consumers are searching for on the web and how consumers do it [2]. Much as Diana Forsythe once argued that "designing and implementing appropriate automated solutions presumes knowledge of physicians' information needs" [3] and pioneered the method of ethnographic techniques to facilitate direct observation of communication about information needs of physicians, we think that understanding consumer health information needs is a prerequisite for building consumer health informatics solutions [4;5] and in turn requires direct observation of information seeking behavior of consumers. Our own previous work in this area includes a semi-quantitative content analysis of emails from patients contacting physicians[6], qualitative research with focus groups, and a direct usability lab observation on how consumers search the web [7]. In the present study we aimed to determine the actual prevalence of health-related searches on the web by analyzing search terms entered by people into popular search engines and to make some preliminary attempts in qualitatively describing and classifying these searches.

Research looking into the prevalence of health-related searches (or related research attempting to determine for example the number of health-related websites) is complicated by the difficulty of defining what "health-related" means. The WHO definition of health as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity" (Preamble to the Constitution of the World Health Organization) is so broad that even financial information could be argued to be "health-related".

The difficulties of defining what is "health-related" (and the time consuming manual coding process) led us to develop and validate a simple algorithm to automatically identify health-related searches. This method also offers

an operational definition of what a "health-related" information or search expression constitutes.

Methods

Harvesting of search queries

The search terms were harvested from two search engines that allow "peeking" of searches, i.e. users can see what queries other users are currently entering. The two search engines used were Metaspy (http://www.metaspy.com/), which lists searches from Metacrawler and AskJeeves (http://www.ask.com/docs/peek/). While Metaspay lists traditional search terms, Askjeeves queries are in the form of questions, for example "Where can I find information on marine plants and algae?". A timed script was developed that periodically visited and "screen scraped" this information, i.e. the HTML was parsed, the appropriate information (the search query) extracted from it and written into a database. 2985 search queries were harvested from MetaSpy between February 2001 and April 2002, and 475 search questions between February 2001 and April 2001. This should constitute a random sample of search queries entered by users.

Human coding of the search queries

A web-based interface was developed to allow human coders to classify the queries. Metaspy searches were classified into "not health related", "somewhat health related", and "clearly health related". The latter two categories were later collapsed for analysis into one category of "health-related" queries. AskJeeves questions were also classified as "not health related" or "health related", and queries from the latter category were also coded with the Ely taxonomy, which was developed to classify information needs of physicians' and which offers a total of 66 different codes [8]. One aspect of this study was to explore to what degree this taxonomy would be useful and applicable to code consumer questions. All queries were coded by two coders independently from each other and inter-observer reliability coefficient was calculated.

Automated method to classify health-related queries

The high interobserver variability in determining what is health-related and the time-consuming manual coding motivated the development and validation of an automatic method to classify search terms and queries as "health related".

Our method for auto-classification proposes to use the assumption that "health-related" search terms should on the resulting webpages co-occur together with the word "health" more often than search terms which are not health related. We use a search engine (google) to determine the number of pages found with the search term AND the word "health", in relation to the number of pages which are found with the search term alone. This proportion – which in the following we call the "co-occurrence rate" (c) - indicates how frequently the search terms occur on the same page with the word "health" and can be seen as a metric for how health related the search query is:

$$c = \frac{pages\ (query\ AND\ health)}{pages\ (query)}$$

If pages (query)=0 then c:= 0.
If c>= ? then query is said to be health-related, otherwise not health-related.

Where

c	:=	co-occurrence rate
Query	:=	search term(s) in question
Pages()	:=	number of hits (pages) retrieved by Google
?	:=	threshold [0,1]

For search queries or terms which are not health related (such as "London"), c should be small, which means that a very low proportion of pages containing the search terms also contain the word "health". In contrast, for health related searches this figure should be closer to 1, indicating a high co-occurrence with the word "health".
If c is greater or equal than threshold ? then the search query can be considered health-related. The optimal threshold ? which divides health related from non-health related search terms was determined empirically as being .35 (see below), i.e. if more than 35% of the pages with the search terms also contain the word health then the search query can be said to be health-related.
For example, if we want to know whether the search term "ovarian cysts" is health-related, we enter this search term into Google and record the number of pages found (59100 hits), then entered the search term "ovarian cysts health" (note that Google uses an implicit AND operator), which elicited 49800 hits, resulting in a co-occurrence rate of 49800/59100 = 84.2%. In contrast, a search query such as "regents park London" results in a co-occurrence rate of only 59600/10800 = 18.1%, and can therefore be ruled out as "not health-related".

To automatically determine c for each of the 2985 search terms from Metaspy we developed a computer script that uses the Google API (http://www.google.com/apis/) to automatically query the Google database for pages con-

taining the search query– first entered on its own, and then in combination with the word health. The script read out the number of hits (pages) found with these two queries, results were written into a database, and the co-occurrence rate for each query was calculated by dividing the figures according to the formula given above. If the original search term has 0 hits (which occurred 113 times), the co-occurrence rate (COR) was set to zero.

Validation of the auto-classification method

We validated the automatic method described above against human coding of search queries by tabulating the co-occurrence rates against a human consensus classification for each search query. We drew a receiver-operating-characteristics (ROC) curve and a precision-recall curve to determine sensitivity (=recall), specificity and precision (=positive predictive value) of this method for different cut-off points ? .

Results

Manual coding of metacrawler search terms

2985 search expressions harvested from Metacrawler were coded by the two authors (GE, CK) independently from each other as "health-related" (including "somewhat health-related) or "non-health related". 108 (3.6%) queries were coded by both coders as "health-related", 2827 (94.7%) of the searches were concordantly classified as non-health related, and 50 (1.7%) received a discordant classification. Search expressions which were coded discordantly included for example "treatment for addiction to porn", "transex", "stop thumb sucking", 'Statistics On Teenage Suicides", "calcium crystals" etc. These queries illustrate the occasional difficulty to determine what "health-related" constitutes. We went through all discordant search expressions again in order to determine a consensus coding. Most of the searches coded by one of the coders as health-related eventually received a consensus coding as "health-related". According to the final consensus rating 135 (4.5%) of all search terms could be considered "health-related". Although not formally coded, most of the remaining search queries appeared to be related to pornographic material.

Manual coding of AskJeeves questions

The 475 AskJeeves questions were coded by two coders independently from each other (Table 1). The two coders identified 48 and 45 (10.1% and 9.5%) health-related questions, respectively. 44 questions (9.3%) were consistently coded as "health-related" (discordantly coded questions include for example "What does every baby need to thrive?", "How long will I live?", "Where can I see pictures of DNA?", and "Where can I find resources from Britannica.com on apathy?"). 36 of these were coded using concordant Ely codes, while 8 questions were coded discordantly. The vast majority of questions (22) were coded as "nonmedical – education – patient". The Ely taxonomy – originally developed to classify physicians' information needs – proved not to be very useful to code consumer questions. We are therefore currently developing a new coding system for consumer health questions.

Table 1. Codings of the AskJeeves questions with the Ely classification [8]

Coder 1	Coder 2	n
Not health related	Not health related	426
Not health related	Epidemiology - not elsewhere classified	1
Diagnosis - cause/ interpretation of clinical finding-symptom		4
Diagnosis - orientation – condition		1
Treatment - drug prescribing - adverse effects - findings caused by drug/ adverse effects of drug		1
Treatment - drug prescribing - orientation/ composition		1
Treatment- drug prescribing - mechanism of action	Treatment - drug prescribing - orientation/ composition	1
Treatment - not limited to but may include drug prescribing - how to do it	Not health related	1
Treatment - not limited to but may include drug prescribing - how to do it		2
Treatment - not elsewhere classified		1
Management (not specifying diagnostic or therapeutic) - not elsewhere classified		1
Management (not specifying diagnostic or therapeutic) - not elsewhere classified	Epidemiology - not elsewhere classified	1
nonclinical - education-patient	Not health related	3
nonclinical - education-patient	Diagnosis - orientation - condition	6
nonclinical - education-patient	treatment - not elsewhere classified	1
nonclinical - education - patient		22
nonclinical - education-patient	nonclinical - not elsewhere classified	1
nonclinical - education-patient	nonclinical - legal	1
		475

During the coding it also became clear that the questions provided by AskJeeves as "what people are asking right

now“ were unlikely to be actual questions asked by people. Rather, the majority of question seemed to be “pre-canned” questions provided by AskJeeves which were mapped to the queries entered by users. Further, the questions seemed to have been filtered, as no sexually oriented questions were displayed (which in the metaspy analysis constituted the majority of searches). Thus, the displayed AskJeeves questions likely provide a biased and non-representative view on the information needs of people. This explains the higher prevalence of health-related searches as compared to the Metaspy searches.

Validation of the automatic classification method

The “co-occurrence rates” for each search query of the metacrawler searches as metrics for their relation to health were calculated as described above (as the ratio between pages with the search term *and* health to pages with the search term alone). Figure 1 shows the distribution of the co-occurrence rates in the entire Metacrawler dataset.

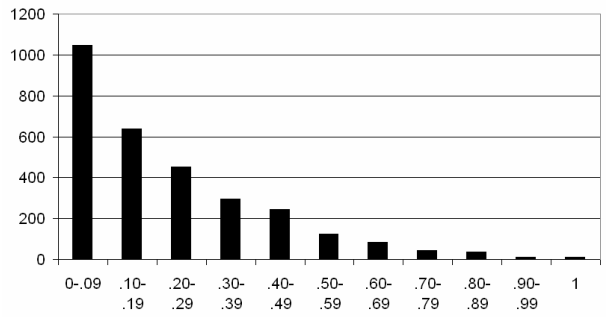

Figure 1. Distribution of “co-occurrence rates” of search terms from metacrawler. The higher the rate, the higher the proportion of pages where the search query and the word health occur together and presumably the more the search query is related to health. The co-occurrence rate can be seen as a “health-relatedness index”.

In order to find the optimal threshold ? (cut-off point for the co-occurrence rate) which discriminates optimally between health-related and non health-related search terms we drew a Receiver Operating Characteristics (ROC) curve with varying threshold parameters ? (Figure 2), showing the test characteristics as a trade off between specificity and sensitivity, with the human codings as a gold standard and the automatic classification with the threshold ? as a test for how health-related a search is. Another way to evaluate the method is to look at the precision-recall curve (Figure 3).

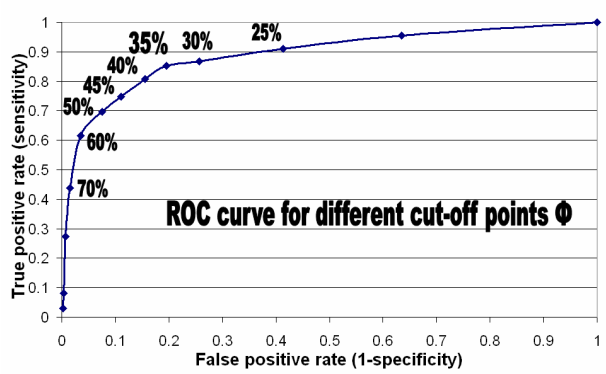

Figure 2. ROC (receiver operating characteristics) curve

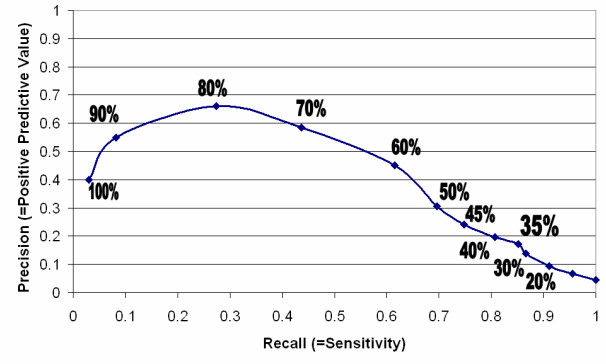

Figure 3. Precision-recall curve

The ROC curve illustrates that a cut-off point of ? =35% has an optimal trade-off between a sensitivity of 85.2% (115/135, ie the proportion of health-related terms correctly picked up by this method) and specificity of 80.4% (2292/2850 not-health related search terms were filtered out) (see Table 2).

The automatic method can be made more sensitive (at the expense of specificity) if a lower threshold ? is chosen. For example if a qualitative researcher wants to do a pre-selection of all possibly health-related search queries without risking to filter out too many true health-related terms, he/she would choose a lower threshold. For example, with a ? of 20% the method still picks up 123/135 of health related searches (a sensitivity of 91.1%), with a specificity of 58.7% (1673/2850 not health related terms are filtered out). Without missing many health-related terms te researcher only has to sift through less than a half (45.6%) of the original search queries.

For other applications a highly specific classify classification might be appropriate. For example, if the threshold is set at 80%, only 56 search queries in total are left, and the method is maximally precise, with a precision (=positive predictive value) of 66% (meaning that 66% of the 56

search terms are in fact health related). The specificity is 99.3%, but the sensitivity (recall) is 27.4% (meaning that only 37/135 health-related terms are in the final set).

Table 2. Contingency table with true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) for the auto-classification method with a threshold ? =.35 for the co-occurance rate

	Manual coding		
Auto-coding	Health-related	Non-health	
c >= .35	115 (TP)	558 (FP)	**673**
c < .35	20 (FN)	2292 (TN)	**2312**
	135	**2850**	2985

Discussion

Based on our analysis we estimate that approximately 4.5% of all searches on the web might be health-related. Although health-related queries constitute a relatively small fraction of web-searches, the absolute numbers are still impressive: Google reports 150 Million searches per day on all regional partner sites combined, which means that about 6.75 Million health related searches *per day* in Google alone are being conducted. In comparison, in 1996 NLM reported 7 Million searches in the MEDLARS (Medline) system *per year*.

While our prevalence estimate of 4.5% is based on data from a single search engine (MetaCrawler), there is little reason to believe that a more commonly used search engine such as Google has a different prevalence of health-related searches. The higher prevalence of health-related searches in AskJeeves is likely a result of a biased (filtered) display of queries on that site.

We think that direct analysis of searches elicit a much more accurate picture of what people are doing and looking for on the web than for example survey data such as the Pew Internet Survey, which currently dominate the literature. Not only is it difficult for people to recall in a survey which kind of information they retrieve on the web most frequently, the accuracy of survey data also suffers from a social desirability bias – rarely people will for example admit to be seeking pornographic material, although these kind of searches are apparently the most prevalent.

To facilitate further research and classification of search queries we also developed and validated an automatic method to identify health-related searches. This method, which looks at co-occurrences of the terms with the word health, can possibly be expanded to classify any kind of short phrases or text as health-related. Potential applications include the automatic analysis of emails and classification into health-related and non-health related, in order to triage incoming emails automatically to technical or medical staff. Each sentence of the email could undergo a co-occurrence analysis and an average co-occurrence rate can be calculated. A validation of this method is currently in progress.

Acknowledgements: Jim Lai programmed the scripts, David Mason provided technical input.

References

1. Pew Internet and American Life Project. The Online Health Care Revolution: How the Web helps Americans take better care of themselves. 11-26-2000.

2. Stavri PZ. Personal health information-seeking: a qualitative review of the literature. *Medinfo* 2001;**10**:1484-8.

3. Forsythe DE, Buchanan BG, Osheroff JA, Miller RA. Expanding the concept of medical information: an observational study of physicians' information needs. *Comput Biomed Res* 1992;**25**:181-200.

4. Eysenbach G. Consumer health informatics. *BMJ* 2000;**320**:1713-6.

5. Houston TK, Chang BL, Brown S, Kukafka R. Consumer health informatics: a consensus description and commentary from American Medical Informatics Association members. *Proc AMIA Symp* 2001;269-73.

6. Eysenbach G,.Diepgen TL. Patients looking for information on the Internet and seeking teleadvice: motivation, expectations, and misconceptions as expressed in e-mails sent to physicians. *Arch Dermatol* 1999;**135**:151-6.

7. Eysenbach G,.Köhler C. How do consumers search for and appraise health information on the World-Wide-Web? Qualitative study using focus groups, usability tests and in-depth interviews. *BMJ* 2002;**324**:573-7.

8. Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA *et al.* A taxonomy of generic clinical questions: classification study. *BMJ* 2000;**321**:429-32.